TITLE: Method and arrangement for avoiding loss of error-critical non real time data during certain handovers

TECHNOLOGICAL FIELD

The invention concerns generally the protocol structures that are used to arrange the communication between a mobile terminal and a packet-switched network. Especially the invention concerns the optimal composition of such structures from the point of view of minimized risk of losing certain types of data in certain handover situations on one hand and reduced complexity on the other.

BACKGROUND OF THE INVENTION

Fig. 1 illustrates the known data protocol stacks that are applied in a packet-switched communication connection where one end is a Mobile Station (MS) and the communication takes place over a GPRS network (General Packet Radio Service) through a Base Station Subsystem (BSS), a Serving GPRS Support Node (SGSN) and a Gateway GPRS Support Node (GGSN). The protocol layers where the peer entities are in the MS and the BSS are the physical layer 101 that employs the GSM cellular radio system (Global System for Mobile telecommunications), the Media Access Control (MAC) layer 102 and the Radio Link Control layer 103 which sometimes is regarded as only a part of the MAC layer 102 - hence the dashed line between them. The protocol layers where the peer entities are in the BSS and the SGSN are the L1bis layer 104, the Network Service layer 105 and the BSS GPRS Protocol (BSSGP) layer 106.

The layers for which the peer entities are in the MS and the SGSN are the Logical Link Control (LLC) layer 107 and the SubNetwork Dependent Convergence Protocol (SNDCP) layer 108. It should be noted that only data or user plane protocols are shown in Fig. 1; a complete illustration of protocols would include the Layer 3 Mobility Management (L3MM) and Short Message Services (SMS) blocks on top of the LLC layer 107 in parallel with the SNDCP layer 108. Additionally there are the known Session Management (SM) and Radio Resource management (RR) entities that are not located on top of the LLC layer. At the interface between the SGSN and the GGSN there are the Layer 1 (L1) layer 109, the Layer 2 (L2) layer 110, a first Internet Protocol (IP) layer 111, the User Datagram Protocol /

Transport Control Protocol (UDP/TCP) layer 112 and the GPRS Tunneling Protocol (GTP) layer 113. Between the MS and the GGSN there are the X.25 layer 114 and a second Internet Protocol layer 115. An application layer 116 in the MS will communicate with a peer entity that is located for example in another MS or some other terminal.

Proposals for the future UMTS (Universal Mobile Telecommunication System) have suggested similar protocol structures for the communication between mobile stations, Radio Network Controllers (RNCs) and service nodes of packet-switched networks, with small changes or modifications in the designations of the devices, layers and protocols. It is typical to protocol structures like that in Fig. 1 that each layer has an exactly determined set of tasks to perform and an exactly determined interface with the next upper layer and the next lower layer. A certain amount of memory and processing power must be allocated in the devices taking part in the communication to maintain the layered structure and accomplish the tasks of each layer. It is therefore easily understood that the more complicated the structure of layered protocols, the more complicated the required software and hardware implementation. Complexity is disadvantageous in terms of costs incurred in design and manufacture and it increases the possibility of design errors. Additionally, in battery-driven mobile terminals it is a continual aim to reduce power consumption and diminish physical size, whereby a more simplified structure of protocol layers would create advantage.

SUMMARY OF THE INVENTION

It is therefore an object of the present invention to provide a method and arrangement that would accomplish the tasks of known communication protocol arrangements but with a simpler protocol structure.

The objects of the invention are accomplished by replacing certain parts of the protocol structure by a temporary suspension of certain communications for the duration of a handover.

The method according to the invention is characterised by that it comprises the steps of

- suspending at least one active non-real time telecommunication connection between a mobile station and the fixed parts of a mobile telecommunication system,

- performing a handover from a first network connection to a second network connection and
- resuming the suspended non-real time telecommunication connection.

The invention also concerns a mobile station arranged to perform a handover according to the above-described method.

The invention relates closely to the observation that the role of certain layers in many protocol structures is of minor practical value and is limited to certain measures for avoiding loss of data during a handover. If the data concerned allows for some additional delays to be caused on its path from the transmitting device to the receiving device, such protocol layers may be omitted altogether by simply suspending the transmission of data when a handover is about to take place and resuming normal operation after the handover has been successfully completed.

In the GPRS example presented in the description of prior art the protocol layer that can be omitted by employing the suspension-resumption mechanism is the LLC layer. We may note that the RLC layer is capable of performing all required error correction tasks over the radio interface in normal operation and the role of LLC has mainly been related to handovers between different BSCs (Base Station Controllers), where error-critical (but not delay-critical) data has needed a mechanism for avoiding loss of data. In the proposed UMTS a similar need has existed in handovers between different RNCs or SGSNs (often designated as 3GSGSNs or 3rd Generation SGSNs). If we remove this need by temporarily suspending the transmission of such error-critical data altogether for the duration of that time interval where loss of data could otherwise occur, the error-correcting functions of the LLC layer become superfluous.

The LLC layer has also had certain responsibilities for flow control. According to the invention the RLC layer may take care of all flow control between the mobile station and a base station controller or a radio network controller (or generally the radio access network), and local flow control mechanisms may be employed for controlling the flow over the interface between the radio access network and a core network. In UMTS the latter is known as the Iu interface.

BRIEF DESCRIPTION OF DRAWINGS

The novel features which are considered as characteristic of the invention are set forth in particular in the appended Claims. The invention itself, however, both as to its construction and its method of operation, together with additional objects and advantages thereof, will be best understood from the following description of specific embodiments when read in connection with the accompanying drawings.

Fig. 1 illustrates the known protocol stacks in a GPRS implementation,

Fig. 2 illustrates the known functional model of an LLC layer,

Fig. 3 illustrates a functional model that would replace the LLC layer according to the invention,

Fig. 4 illustrates an arrangement of protocol stacks according to the invention,

Figs. 5a to 5c illustrate an inter-RNC, intra-SGSN handover according to the invention,

Figs. 6a to 6c illustrate an inter-RNC, inter-SGSN handover according to the invention and

Figs. 7a and 7b show a comparison between a prior art method and a method according to the invention.

DETAILED DESCRIPTION OF THE INVENTION

We will illustrate the applicability of the invention in connection with the known GPRS system. However, the presented examplary embodiments do not limit the applicability of the invention to any specific system. As a background to the invention we will first consider some known characteristics of the GPRS system.

The general packet radio service (GPRS) is a new service to the GSM system, and is one of the objects of the standardization work of the GSM phase 2+ at the ETSI (European Telecommunications Standards Institute). The GPRS operational environment comprises one or more subnetwork service areas, which are inter-connected by a GPRS backbone network. A subnetwork comprises a number of packet data service nodes (SN), which in this application will be referred to as serving GPRS support nodes (SGSN), each of which is connected to the mobile

telecommunications system in such a way that it can provide a packet service for mobile data terminals via several base stations, i.e. cells. The intermediate mobile communication network provides packet-switched data transmission between a support node and mobile data terminals. Different subnetworks are in turn connected to an external data network, e.g. to a public switched data network (PSDN), via GPRS gateway support nodes (GGSN). The GPRS service thus allows to provide packet data transmission between mobile data terminals and external data networks when the appropriate parts of a mobile telecommunications system function as an access network.

In order to access the GPRS services, a MS shall first make its presence known to the network by performing a GPRS attach. This operation makes the MS available for SMS (Short Message Services) over GPRS, paging via SGSN, and notification of incoming GPRS data. More particularly, when the MS attaches to the GPRS network, i.e. in a GPRS attach procedure, the SGSN creates a mobility management context (MM context). Also the authentication of the user is carried out by the SGSN in the GPRS attach procedure. In order to send and receive GPRS data, the MS shall activate the packet data address that it wants to use, by requesting a PDP context activation procedure, where PDP comes from Packet Data Protocol. This operation makes the MS known in the corresponding GGSN, and interworking with external data networks can commence. More particularly a PDP context is created in the MS and the GGSN and the SGSN. The PDP context defines different data transmission parameters, such as the PDP type (e.g. X.25 or IP), PDP address (e.g. X.121 address), quality of service (QoS) and NSAPI (Network Service Access Point Identifier). The MS activates the PDP context with a specific message, Activate PDP Context Request, in which it gives information on the TLLI, PDP type, PDP address, required QoS and NSAPI, and optionally the access point name (APN).

The quality of service defines how the packet data units (PDUs) are handled during the transmission through the GPRS network. For example, the quality of service levels defined for the PDP addresses control the order of transmission, buffering (the PDU queues) and discarding of the PDUs in the SGSN and the GGSN, especially in a congestion situation. Therefore, different quality of service levels will present different end-to-end delays, bit rates and numbers of lost PDUs, for example, for the end users.

Currently the GPRS allows for only one QoS for each PDP context. Typically a terminal has only one IP address, so conventionally it may request for only one PDP context. There is recognised the need for modifying the existing systems so that a

PDP context could accommodate several different QoS flows. For example, some flows may be associated with E-mail that can tolerate lengthy response times. Other applications cannot tolerate delay and demand a very high level of throughput, interactive applications being one example. These different requirements are reflected in the QoS. Intolerance to delay must usually be associated with a relatively good tolerance for errors; correspondingly a very error-critical application must allow for long delays, because it is impossible to predict how many retransmission attempts it will take to achieve the required high level of correctness. If a QoS requirement is beyond the capabilities of a PLMN, the PLMN negotiates the QoS as close as possible to the requested QoS. The MS either accepts the negotiated QoS, or deactivates the PDP context.

Current GPRS QoS profile contains five parameters: service precedence, delay class, reliability, and mean and peak bit rates. Service precedence defines some kind of priority for the packets belonging to a certain PDP context. Delay class defines mean and maximum delays for the transfer of each data packet belonging to that context. Reliability in turn specifies whether acknowledged or unacknowledged services will be used at LLC (Logical Link Control) and RLC (Radio Link Control) layers. In addition, it specifies whether protected mode should be used in case of unacknowledged service, and whether the GPRS backbone should use TCP or UDP to transfer data packets belonging to the PDP context. Furthermore, these varying QoS parameters are mapped to four QoS levels available at the LLC layer.

Fig. 2 is a functional model of a known LLC protocol layer 201, corresponding to the blocks 107 in Fig. 1. Block 202 represents the known lower layer (RLC/MAC; Radio Link Control / Media Access Control) functions that are located below the LLC layer 201 in the protocol stack of a mobile station MS. Correspondingly block 203 represents the known lower layer (BSSGP) functions that are located below the LLC layer 201 in a serving GPRS support node SGSN. The interface between the LLC layer 201 and the RLC/MAC layers 202 is called the RR interface and the interface between the LLC layer 201 and the BSSGP layers 203 is called the BSSGP interface.

Above the LLC layer there are the known GPRS Mobility Management functions 204 (also known as the Layer 3 Mobility Management functions or L3MM), SNDCP functions 205 and Short Messages Services functions 206. Each one of these blocks has one or more interfaces with the LLC layer 201, connecting to its different parts. The Logical Link Management Entity 207 has an LLGMM control interface (Logical Link - GPRS Mobility Management) with block 204. Mobility

management data is routed through a LLGMM data interface between block 204 and the first Logical Link Entity 208 of the LLC layer. The second 209, third 210, fourth 211 and fifth 212 Logical Link Entities connect to block 205 through the corresponding interfaces; according to the QoS levels handled by each of the Logical Link Entities the interfaces are known as QoS 1, QoS 2, QoS 3 and QoS 4. The sixth Logical Link Entity 213 of the LLC layer connects to block 206 via an LLSMS interface (Logical Link - Short Messages Services). The Service Access Point Identifiers or SAPIs of the first 208, second 209, third 210, fourth 211, fifth 212 and sixth 213 Logical Link Entities are respectively 1, 3, 5, 9, 11 and 7. Each one of them is connected inside the LLC layer to a multiplexing block 214, which handles the connections through the RR interface to block 202 and further towards the mobile station as well as the connections through the BSSGP interface to block 203 and further towards the SGSN. The connection between the multiplexing block 214 and the lower layer block 202 in the direction of the MS may be described as a "transmission pipe".

Fig. 3 illustrates an arrangement according to the invention where the LLC layer has been completely omitted. The upper layers comprise a MM/RR part 301 for known mobility and radio resource management, an SMS part 303 for processing data related to short messages, as well as a part 302' for processing the received data and data to be transmitted according to other functionalities. "Local" multiplexing/-demultiplexing is performed at the upper layers in blocks 304 to 308 so that there is only one transmission pipe for control information between the MM/RR part 301 and the lower layers, one transmission pipe for SMS-related information between the SMS part 303 and the lower layers, and one transmission pipe for each quality of service class between the other functionalities part 302 and the lower layers. Multiplexing is shown in Fig. 3 as taking place in separate functional blocks; however, it may be an inherent part of for example one or several functionalities in the other functionalities part 302.

The RLC/MAC layer is located directly under the upper layers in Fig. 3. It performs the known RLC/MAC functions for each flow of information for which there is a connection between it and the upper layers. The MAC functions consist of procedures for sharing the common radio channels between mobile stations as well as allocations and disallocations of dedicated radio channels. The RLC functions comprise the composing and decomposing of RLC blocks, detecting corrupted RLC blocks and arranging for the retransmission of corrupted blocks when appropriate. In UMTS the the concept of an RLC unit is unidirectional and reserved for one

information flow only, so the widely interpreted RLC layer in the protocol structure will accommodate a pair of RLC units for each active flow of information. The multiplexing and demultplexing of the RLC blocks belonging to different flows of information takes place on the physical layer, which is represented by block 315 in Fig. 3. In a spread spectrum system it is advantageous to multiplex all flows of information related to a certain mobile terminal onto a single code channel. From the published standardisation work of the UMTS there is known a physical layer that is applicable to perform the operations represented by block 315.

Fig. 3 as such is only applicable to the mobile station, because there is an RLC / MAC layer under the higher-order layers. However, it is easy to generalise the arrangement of Fig. 3 so that there may be a BSSGP layer under the higher-order layers, resulting in an arrangement applicable to a SGSN. Also in that case there must be an additional stage of multiplexing/demultiplexing at the physical level, like block 315 in Fig. 3.

Fig. 4 illustrates the inventive structure of protocol stacks which is comparable to the known arrangement of Fig. 1. It is noted that there is no LLC layers in the mobile station or the SGSN, the physical layer between the mobile station and the RAN has been replaced by a UMTS physical layer 401, the BSSGP layer between the RAN and the SGSN has been replaced by a corresponding UMTS layer preliminarily known as the RANGP (RAN GPRS Protocol) layer 402, and the MAC, RLC, SNDCP, Network Service and L1bis layers have been adapted according to the guidelines given above in association with Fig. 3.

Next we will describe some handover situations where a mobile station and the network will apply the principle of temporarily suspending error-critical communications according to the invention. Fig. 5a illustrates a situation where the mobile station 501 has a macrodiversity connection with two RNCs (Radio Network Controllers) network so that the first RNC 502 is the so-called serving RNC and the second RNC 503 is the so-called drifting RNC. The interface between the two RNCs is called the Iur interface. From the serving RNC 502 there is a connection to a SGSN 504 over a so-called Iu interface, and from the SGSN there is a connection to a GGSN 505. A generalisation of the arrangement of Fig. 5a is the case where the second RNC is just a "new" serving RNC regardless of whether it was firstly a drifting RNC or not. Drifting RNCs relate only to macrodiversity; if no macro-diversity is applied there will be an "old" serving RNC and a "new" serving RNC (or, in second generation systems, an "old" BSS and a "new" BSS) with little simultaneous service from both of them to the mobile station.

In Fig. 5b either the mobile station 501 or some network device in the radio access network (not shown) where the serving RNC 502 is located notices that the direct connection between the mobile station and the serving RNC is critically weakening or has been severed, so a handover to the second RNC 503 is inevitable. According to the invention, the handover is started by requesting all such active services to be suspended which require a high level of correctness and tolerate long delays. In a GPRS type arrangement the suspension of services would require suspending whole PDP contexts, because the PDP context can only have one QoS. In a UMTS type arrangement it suffices to suspends those QoS flows for which the QoS allows for (delay tolerance) and even requires (correctness) the suspension. To retain generality we will use the word "service" for the entities that will be suspended. It will be most advantageous to define beforehand, in a standardised specification, a threshold value either for required correctness or for allowed delays or for both so that only those active services will be suspended for which the required correctness or allowed delay or both exceed the threshold value(s).

After the suspension of the selected active services the network will establish a new connection over the Iu interface between the second RNC 503 and the SGSN 504. Simultaneously communication on the non-suspended services may continue. Typically there will be some RLC level buffers in at least one of the devices taking part in the communication that need to be emptied before the second RNC may be designated as the serving RNC. The situation illustrated in Fig. 5c may only become relevant after all such RLC buffers have been emptied and the new connection over the Iu interface between the second RNC and the SGSN has been established. At that moment the suspended services may be released so that communication over them will continue normally. In Fig. 5c the second RNC 503 is the serving RNC and the old connection over the Iu interface between the first RNC 502 and the SGSN 504 has been terminated.

In Fig. 5c it has been assumed that the handover was not associated with a complete severance of the direct connections between the mobile station 501 and the first RNC 502. Consequently the connection over the Iur interface between the RNCs is not terminated and the first RNC continues to operate as a drifting RNC. Sooner or later, especially if the mobile station continues the movement that caused the direct connections to the first RNC to weaken, these direct connections will fall under the level of acceptable quality so that they are completely released and the connection over the Iur interface between the RNCs is terminated.

Figs 6a to 6c describe a handover situation where the new RNC operates under a new SGSN. Such a handover is called an inter-RNC, inter-SGSN handover. Here we have expected that an Iur interface exists even between RNCs that operate under different SGSNs; this is not a requirement associated with the invention, because the invention works equally well without any connections between the RNCs. Fig. 6a corresponds to Fig. 5a with the sole exception that the second RNC 601 belongs to the domain of a second SGSN 602. At some stage it is again noticed that a handover from the first RNC to the second RNC will be required. The operation starts with the suspension of error-critical, delay-tolerant PDP contexts as described above. According to Fig. 6b, the controlling responsibility remains in the first RNC 502 and the first SGSN 504 during the time it takes the mobile station to registrate under the second SGSN 602 and the latter to set up a new GTP bearer with the GGSN 505. The first SGSN 504 will also transmit all information related to the connections to be transferred to the second SGSN 602, as illustrated by an arrow in Fig. 6b. Thereafter the controlling responsibilities may be moved to the second RNC 601 and the second SGSN 602 as illustrated in Fig. 6c, and the suspended PDP contexts may be resumed. If there are still usable direct connections between the mobile station and the first RNC 502 and a working Iur interface between the RNCs, the first RNC may remain as a drifting RNC.

It is possible that the new SGSN is not capable of handling some information flows the controlling responsibility of which it has received during the handover. Special measures which are as such outside the scope of the present invention may be taken in order to adapt the information flows to the capabilities of the new SGSN. After all information flows are in such shape that the new SGSN is capable of handling them, the connection between the mobile terminal and the old SGSN may be terminated.

Figs. 7a and 7b are simplified flow diagrams that show an important difference between a prior art method (Fig. 7a) and a method according to the invention (Fig. 7b) when handovers are concerned. In Fig. 7a all QoS flows are active throughout the handover, and LLC layer routines are employed to correct any errors that the handover causes to the error-critical, delay-tolerant QoS flows. In Fig. 7b a step 704 of suspending the selected error-critical, delay-tolerant QoS flows precedes the handover, no LLC layer routines are performed, and a step 705 of resuming the error-critical, delay-tolerant QoS flows follows the handover.

A comparison between Figs. 1 and 4, with the help of Figs. 2 and 3, may be used to describe a mobile station and an SGSN according to the invention. It is known as

such that the advantageous implementation of the protocol stacks in mobile stations and SGSNs is in the form of microprocessor-executable computer programs stored in memory devices. By applying the teachings of the present patent application it is within the capabilities of a person skilled in the art to realise, instead of the protocol structures illustrated in Figs. 1 and 2, the protocol structures according to Figs. 3 and 4 so that the mobile stations and SGSNs with such an implementation will operate according to the present invention.

The invention has been desribed above solely with reference to packet-switched non-real time communication connections. However, it is possible to apply the concept of suspending and releasing also to specific kinds of circuit-switched connections. The prerequisite for applying the invention to circuit-switched connections is that such connections must have very relaxed delay requirements; in the terminology of second generation digital cellular radio systems the invention is applicable to non-transparent circuit-switched connections but not to transparent circuit-switched connections because of the tight delay requirements associated therewith.